Matching References with MEDLINE via TCP/IP

John N. Guidi, The Jackson Laboratory, Bar Harbor, Maine

Bibliographic references are an important part of databases and information resources used by clinicians and biomedical researchers. In addition to the obvious clerical advantages of standard references, bibliographic references can also be used as links to related items in different data sets. This paper describes an effort that involved matching references from a variety of disparate databases to canonical MEDLINE references. The references matched were those involved in a database unification project which is part of the Mouse Genome Informatics effort at The Jackson Laboratory. Software was developed to take advantage of a commercially available retrieval engine which accesses MEDLINE on CD-ROM disks. The software permits client programs on UNIX/C, and potentially other environments, to access unabridged MEDLINE via networks supporting the TCP/IP protocols. The matching process described can be used as a model for similar efforts with different research or clinical data sets, as well as different hardware or software environments.

INTRODUCTION

There are obvious clerical advantages in eliminating duplicates, providing standardized presentations, and generally improving the quality and content of bibliographic references in databases used by clinicians and biomedical researchers. In addition, bibliographic references can be used as a key to identify potential relationships among data, and can be used to link separate databases together, often in unanticipated ways [13].

One of the goals of the Mouse Genome Informatics effort at The Jackson Laboratory, supported by the Human Genome Office, is to unify a number of existing disparate databases, and fold them into a single comprehensive database, the Mouse Genome Database (MGD). Of consideration here, there are 8 resources of information that have different origins and that are supported by different hardware and software systems. These 8 databases form the foundation of MGD, and all of the data described in these databases have references to the primary literature. MEDLINE provides a source of references to the biomedical literature that can be used as the standard to which our corresponding references can adhere [7]. The vast majority of references under consideration are available from MEDLINE, although there is also a requirement to include references outside of MEDLINE coverage (e.g., references from books, meeting abstracts, journals not covered by MEDLINE, etc.). This paper describes the efforts to provide a single source of references (i.e., a master bibliography) for the comprehensive database MGD, and in particular, illustrates how references can be matched to canonical references available from MEDLINE.

Each of the databases under consideration provides references in a format peculiar to that individual database. In most cases, the reference is simply unstructured text. There are many typographical, spelling, and other errors. In some databases, the same reference is entered a number of times in alternative forms. The origins and evolution of these resources were such that there was minimal coordination of effort. For each reference from the originating databases, the goal is to attempt to find the corresponding MEDLINE reference. If found, the appropriate MEDLINE unique identifier (UI) and other fields available from the MEDLINE unit record supersede the original reference. Figure 1 illustrates a canonical MEDLINE reference followed by samples of variants of this reference available from the originating databases.

AU - Deol MS
TI - The neural crest and the acoustic ganglion.
SO - J Embryol Exp Morphol 1967 Jun;17(3):533-41

TI - The neural crest and the acoustic ganglion.
SO - J. Embryol. Exp. Morphol. 17:535-541.

AU - Deol MS
TI - The neural crest and __ acoustic ganglion.
SO - J. Embryol. Exp. Morphol. 17:533-541.

AU - Deol MS
TI - The neural crest and the acoustic ganglion.
SO - J Embryol Exptl Morphol 1967; 17:533-41

Figure 1: Example Variant References

The Knowledge Host product, from Aries Systems Corporation, was used as a retrieval engine to provide access to MEDLINE on CD-ROM.

0195-4210/92/$5.00

Software was written to extend the query interface to compliant programs on networks supporting the TCP/IP protocols. References from the originating databases were parsed into fields that were then used by a number of programs that attempted to find the matching canonical MEDLINE references. 92% of 14,701 references from the originating databases were matched to corresponding MEDLINE references using these methods.

MEDLINE RETRIEVAL ENGINE

The Knowledge Host product was used as a source of unabridged MEDLINE references for this effort [1]. It is a Macintosh background application that serves MEDLINE references to other Macintosh applications via a file based protocol. Unabridged MEDLINE includes coverage from 1966 to present on 7 compact disks (CD-ROMs) that contain all the compressed data and indices broken into discrete years. As with any CD-ROM based coverage, there is a gap between what is currently available from online MEDLINE and the latest CD-ROM. With respect to this particular query engine, there are some constraints as to the questions that can be efficiently resolved. The maximum number of documents that can be returned for any single query is restricted to 1,000. Knowledge Host resolves queries using a combination of indexing techniques and probabilistic information retrieval techniques.

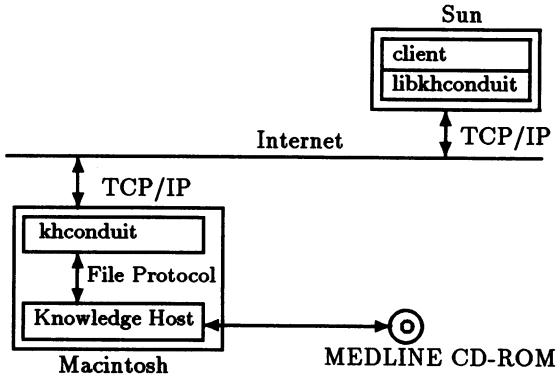
Indices of use to the present effort include: Author Name, Journal Title Abbreviation, Unique Identifier. Exact matches of terms in indices is supported, as well as tail truncation matches. For example, a query for the journal term "Age#" would include the journals "Agents Action", "Age Ageing", etc. There is no support for an index cursor, thus it is not possible to scan an index. Although a Publication Date index is available, it was not useful. This index contained terms of the form "YYYY MMM" where YYYY is the year and MMM is the month. Few of the references from the originating databases had any additional information regarding publication date other than year. The computational expense for Knowledge Host to tail truncate a year (e.g., 1996#) to include all the months (i.e., Jan, Feb, etc.) is high. The fact that there are only a relatively small number of discrete years covered on any single CD-ROM turns out to be useful enough in practice as an alternative to index access with tail truncation for dates.

Probabilistic methods are used to provide support for free form topic statement searches [10]. Potential documents satisfying free form term queries are collected and weighed with respect to the likelihood that they indeed satisfy the request. A user supplied threshold indicates the limit of the number of documents returned. Free form document searches can be computationally expensive, and their running times are dependent on the frequency of occurrence (i.e., popularity) of the terms in question in MEDLINE.

MEDLINE VIA TCP/IP

The target for the comprehensive database MGD is a UNIX environment. At the time this effort was undertaken, there were no systems or products available that would permit application programs on UNIX systems to query MEDLINE via a program interface. One contribution that this paper documents is the software distribution that consists of a Macintosh application (khconduit), a UNIX/C library for compliant khconduit clients (libkhconduit), and a number of UNIX/C khconduit clients. Figure 2 illustrates the architecture of this simple client/server relationship [4]. These software components, all produced at The Jackson Laboratory, leverage the functionality of the commercial product Knowledge Host, by providing UNIX applications the ability to query a source of unabridged MEDLINE references via a program interface to an appropriately configured Macintosh. The network requirement is that the Macintosh support the TCP/IP protocols [8, 9].

Figure 2: Client/Server MEDLINE Architecture



The Macintosh application khconduit acts as a conduit, hence the name, between the Knowledge Host background application running on the same Macintosh and client applications running elsewhere on the network. On the one side, khconduit

uses the disk based protocol required to communicate with Knowledge Host. On the other side, khconduit listens on a TCP port for client connection requests. Client query requests are passed directly to the Knowledge Host retrieval engine to be resolved. Replies from Knowledge Host are passed directly to the client. The current implementation of khconduit limits the number of connected clients to one.

The library libkhconduit packages together the handshaking required with khconduit and provides a program interface to the Knowledge Host retrieval engine. The interface is identical to that provided by Knowledge Host to Macintosh applications, but with the addition of TCP/IP support. This gives clients the ability to directly query unabridged MEDLINE over a TCP/IP network. The current implementation of libkhconduit, which is written for UNIX/C, permits a single client to access multiple khconduit servers, but only in a synchronous manner.

Table 1 illustrates representative response times for various types of artificial queries. Knowledge Host and khconduit were running on a Macintosh IIci with 8 MB of memory, a direct ethernet connection, and an external CD-ROM drive. The libkhconduit library and throughput clients were running on a Sun SPARCstation 2 with 16MB of memory. In each case, the client recorded the actual time between when the first query was made and when the last reply was received. Query Q_UI consists of requests for 100 random unique identifiers known to exist in MEDLINE. Q_AU requests 100 random author names known to exist in MEDLINE and for each author, returns one MEDLINE reference. Q_TA is a predicate test identical to Q_AU, except 100 random journal abbreviations known to exist are used. Query Q_TI takes 100 random references known to exist in MEDLINE and for each reference requests a free form search with all terms in the title. Q_TI_TA takes 100 random references and requests, for each, a search that includes all terms in the title, as well as the associated journal abbreviation.

Queries Q_UI, Q_AU, and Q_TA are resolved solely by consulting the appropriate index. Q_TI involves a probabilistic search which allows one to focus on a narrow set of documents of interest, but at a much larger computational expense. Query Q_TI_TA illustrates that where probabilistic and indexed queries are combined, the probabilistic

Table 1: Response Times (n = 5)

Query	average sec/reference
Q_UI	4.12 ± .07
Q_AU	4.80 ± .13
Q_TA	4.85 ± .15
Q_TI	148.27 ± 34.21
Q_TI_TA	139.06 ± 25.38

processing dominates. These results need to be be considered in determining whether to have the matching clients strive for increased precision versus recall. High precision can be obtained, but at a computationally expensive price. In comparison, high recall, low precision results can be obtained relatively quickly, placing the burden on the client to sift through results to find probable matches.

Although it is possible to have a single Macintosh support multiple CD-ROM disks, it is clear there is an advantage to running multiple Knowledge Host retrieval engines concurrently instead. The fact that discrete time periods of MEDLINE coverage are on separate CD-ROM disks can be used to advantage by having the clients restrict queries by time periods. During the actual matching phase, it was common to have a system topology that included multiple Macintoshes, with disks covering different time intervals, concurrently serving separate processes on a single Sun computer.

PARSING ORIGINAL REFERENCES

The references from the originating databases need to be broken into constituent pieces in preparation for matching [3]. Each set of references needs to be treated individually to accommodate the fact that the references are stored in different software systems with different views of a reference. Although the bulk of the references are unstructured text, within a given database the format tends to be fairly consistent.

Individual YACC grammars were written for 5 of the data sets to parse the unstructured text into fields [11]. The fields chosen are a combination of those permitted in queries with the MEDLINE query engine used and those that can be used to determine if a potential reference is indeed a match. A successful parse provided a reference with the following fields (fields in capital letters are available in the MEDLINE unit record): AU (author), primary (primary author), TI (title), SO

(source), TA (journal title abbreviation), DP (date of publication), year (year of publication), VI (volume issue), IP (issue/part/supplement), PG (pagination), firstpg (first page). A successful parse only indicates that at a syntactic level, there is a reasonable expectation that the fields parsed are in good enough shape to be used in the matching process. Table 2 illustrates the results of parsing. As references from the resources hmdp, mldp, mp, and probes were available in delimited fields, they did not require YACC grammars for parsing.

Table 2: Parsing Results

resource	parsed	unparsed	% unparsed
gbase	3,646	302	7.65
humanref	282	21	6.93
matrix	372	80	17.70
mlc	4,790	279	5.50
hmdp	2,652	-	-
mldp	2,048	-	-
mp	2,410	-	-
probes	1,641	-	-
Totals	17,841	682	

MATCHING MEDLINE

After the references from the original databases were parsed, the effort turned to finding matching MEDLINE references. The problem of matching references to a canonical form can be considered a variation of the problem of elimination of duplicates, where a unique canonical record is included in each matching set. Many previous efforts have used the technique of creating a key for each reference in a data set, with the philosophy that identical matches will have the same key. Hickey and Rypka discuss the effects of a variable length key that averages 52 bytes used to eliminate duplicates in online bibliographic catalogs [5]. Slach discusses tagging each reference with a key constructed with year, first four characters of the first authors last name, and the first page [12]. Garfield discusses use of a 14 character key constructed with the first four characters of the first authors last name, the year of publication, volume, and first page [2]. Yannakoudakis et al. discuss use of the Universal Standard Bibliographic Code (USBC) in matching references between non-standardized databases [15]. The USBC is an 18 byte key that is generated from title, author, year, volume, edition, and pagination - with the title and author being the prime elements. Toney describes algorithms for grouping duplicates and discusses the importance of enumerating classes of errors and the requirement for human intervention [14].

Although it was not practical to construct a key for every MEDLINE reference available with the retrieval engine used, these previous efforts showed the effectiveness of combining specific fields as search criteria. A number of variants of a khconduit compliant client (match_aries) were written to take advantage of the various query methods available by Knowledge Host and the various fields whose importance was illustrated in previous works. Table 3 summarizes the various query strategies used (E = exact match, T = tail truncation, P = probabilistic match). The year is always searched for by virtue of the fact that disks provide coverage for discrete years. Different scoring methods were used to determine a match, depending on the particular query strategy used.

Table 3: Query Strategies for Matching

TA	TI	primary	year
E	-	E	√
E	-	T	√
T	-	-	√
T	P	-	√

RESULTS AND DISCUSSION

The procedures described were used to attempt to find matching MEDLINE references with the 17,841 references parsed from 8 databases. Of these, 13,490 were found to match 8,426 unique MEDLINE references (the ratio of duplication is 1.6:1). There were 3,140 references, which reduced to 1,882 after duplicates were identified, that were known a priori not to exist in MEDLINE. Reasons include publication prior to MEDLINE coverage, journals that MEDLINE does not index, items that MEDLINE does not cover (e.g., books, meeting abstracts, proceedings), etc. There were 1,211 references that were unmatched, which reduced to 1,061 after duplicates were identified, whose disposition is unknown. This set of 1,061 references will need to be inspected manually to determine those references that do indeed have MEDLINE matches that were missed in the matching process, and those that are actually not available in MEDLINE. Excluding references known not to be in MEDLINE, the percentage of database references matched to MEDLINE is 92% (=13,490/(17,841-3,140)).

Originally, it was expected that a complicated process akin to correcting words in text would need to be employed to match the database references [6]. This turned out not to be the case. The percentage of matching (92%), using an existing product which provided a combination of index and probabilistic search of MEDLINE, was quite acceptable. The effort involved in manually matching the relatively few unmatched references is estimated to be small in comparison with the effort that would be required to provide software to handle the number of cases remaining. A side effect of this entire effort is the availability of unabridged MEDLINE on CD-ROM to clients on TCP/IP networks.

ACKNOWLEDGEMENTS

This work has been supported by The Pew Memorial Trust and NIH grants CA34196 and HG00330. Larry Mobraaten of The Jackson Laboratory recognized the importance of standardizing references and secured initial funding. Lyndon Holmes of Aries Systems Corporation provided technical support on the inner workings and specifics of the Knowledge Host retrieval engine. Carolyn Tolstoshev of the National Library of Medicine provided a journal file that was used to bootstrap our canonical journal list. I wish to thank Alex Smith, Carolyn Blake, Susan Dewey, Les Kozak, and Moyha Lennon-Pierce for their generosity and patience, as they provided the environment and computer resources necessary to complete this work. The efforts and results in this paper have been subsumed by the Mouse Genome Informatics Program Project (NIH HG00330) at The Jackson Laboratory and form the basis of the Master Bibliography for this Project.

Reference

[1] "Knowledge Host Background Database Server Rev 2.223." Aries Systems Corporation, 200 Sutton Street North Andover, MA 01845, 1992.

[2] E. Garfield "Journal Editors Awaken to the Impact of Citation Errors. How We Control Them at ISI." *Current Contents*, (41):5-13, October 8 1990.

[3] C. M. Goldstein and M. Prettyman. "Processing Downloaded Citations," in *Downloading/Uploading Online Databases & Catalogs. Proceedings of the Congress for Librarians. February 18, 1985. St. John's University, Jamaica, NY.* Ed. B. H. Weinberg and J. A. Benson. Pierian Press, 1985, pp. 40-48.

[4] J. Gray and A. Reuter. "The Transaction-Oriented Computing Style," in Transaction Processing: Concepts and Techniques. Morgan Kaufmann, 1993. Chap. 5.2.1. pp. 241-249.

[5] T. B. Hickey and D. J. Rypka. "Automatic Detection of Duplicate Monographic Records." *Journal of Library Automation*, 12(2):125-142, June 1979.

[6] K. Kukich "Techniques for Automatically Correcting Words in Text." *ACM Computing Surveys*, 24(4):377-439, December 1992.

[7] National Library of Medicine "Online Services Reference Manual." National Library of Medicine, 1992.

[8] J. Postel "RFC 791: Internet Protocol." DARPA Internet Program Protocol Specification. Information Sciences Institute, University of Southern California. September 1981.

[9] J. Postel "RFC 793: Transmission Control Protocol." DARPA Internet Program Protocol Specification. Information Sciences Institute, University of Southern California. September 1981.

[10] G. Salton "Developments in Automatic Text Retrieval." *Science*, 253:974-980, August 30 1991.

[11] A. T. Schreiner and H. G. Freidman. "Introduction to Compiler Construction with UNIX." Prentice-Hall, 1985.

[12] J. E. Slach "Detection and Elimination of Duplicates from Multidatabase Searches." *Bulletin of Medical Library Association*, 73(3):235-237, July 1985.

[13] W. D. Sperzel, R. M. Abaranel, S. J. Nelson, M. S. Erlbaum, D. D. Sheretz, M. S. Tuttle, N. E. Olson and L. F. Fuller. "Biomedical Database Inter-Connectivity: An Experiment Linking MIM, GENBANK, and META-1 via MEDLINE.," in *Proceedings of the Annual Symposium on Computer Applications in Medical Care.* 1991, pp. 190-193.

[14] S. R. Toney "Cleanup and Deduplication of an International Bibliographic Database." *Information Technology and Libraries*, 11(1):19-28, March 1992.

[15] E. J. Yannakoudakis, F. H. Ayres and J. A. W. Huggill. "Matching of Citations between Non-Standardized Databases." *Journal of the American Society for Information Science*, 41(8):599-610, 1990.